PERCEPTUALLY BASED SPEECH ENHANCEMENT USING THE WEIGHTED β-SA ESTIMATOR

Eric Plourde and Benoît Champagne

McGill University
Department of Electrical and Computer Engineering
Montreal, Quebec, Canada, H3A 2A7
e-mail: eric.plourde@mail.mcgill.ca, benoit.champagne@mcgill.ca

ABSTRACT

In this paper, we first propose a new family of Bayesian estimators for speech enhancement where the cost function includes both a power law and a weighting factor. Secondly, we set the parameters of the estimator based on perceptual considerations by taking into account the masking properties of the ear and the perceived loudness of sound. Our results show that the new estimator achieves better overall performance than existing Bayesian estimators both in terms of objective and subjective measures. Specifically, it shows a segmental SNR improvement of up to 0.65 dB while it obtains the best scores in a MUSHRA test for both white and aircraft cockpit noises.

Index Terms— Speech enhancement, minimum mean square error methods

1. INTRODUCTION

The main objective of speech enhancement techniques is to remove a certain amount of noise from a noisy speech signal while keeping the speech component as undistorted as possible. In the Bayesian approach for speech enhancement, an estimate of the clean speech is derived by minimizing the statistical expectation of an appropriate cost function. A well-known Bayesian estimator is the minimum mean square error (MMSE) estimator of the short-time spectral amplitude (STSA) where the chosen cost function involves the squared difference between the estimated and actual clean speech STSA [1]. The MMSE STSA cost function was recently generalized in two different ways by You *et al.* in [2] and Loizou in [3]. In the β-Order STSA MMSE estimator [2] (which we will denote as β-SA for convenience) a power law (i.e. an exponent β) was applied to the estimated and real clean speech STSA in the squared difference of the cost function. In [3], the squared difference in the MMSE STSA cost function was weighted by the STSA of the clean speech raised to an exponent p; the resulting estimator was termed the Weighted Euclidien (WE) estimator.

Building on the work by You *et al.* [2] and Loizou [3], we first propose a new family of estimators where the cost function includes both a power law and a weighting factor which we call the Weighted β-SA estimator (Wβ-SA). Secondly, we choose the parameter values defining the Wβ-SA estimator (i.e. β and p) based on perceptual considerations by, first, taking into account the masking properties of the ear and, second, considering the perceived loudness of sound instead of its intensity. We find through both objective and subjective experimental measures that the new Wβ-SA estimator, with the values of p and β chosen according to the proposed perceptual approach, shows improvements over the other Bayesian STSA estimators compared (i.e. [1, 3, 4]).

The paper is organized as follows. Section 2 reviews existing Bayesian STSA estimators while Section 3 derives the new Wβ-SA estimator. In section 4, we propose some perceptually relevant values for the parameters of the Wβ-SA estimator (i.e. β and p). Section 5 presents experimental results and related discussions. A brief conclusion follows in Section 6.

2. BAYESIAN STSA ESTIMATORS

Let the observed noisy speech be

$$y(t) = x(t) + n(t), \qquad 0 \leq t \leq T \tag{1}$$

where $x(t)$ is the clean speech, $n(t)$ is the additive noise and $[0, T]$ is the observation interval. Let Y_k, X_k and N_k denote the k^{th} complex spectral components of the noisy speech, clean speech and noise respectively.

In Bayesian STSA estimation for speech enhancement, the goal is to obtain the estimator $\hat{\mathcal{X}}_k$ of $\mathcal{X}_k \triangleq |X_k|$ which minimizes $E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\}$ where $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$ is a chosen cost function and E denotes statistical expectation. This estimator is then combined with the phase of the noisy speech, $\measuredangle Y_k$, to yield the estimator of the complex spectral component of the clean speech $\hat{X}_k = \hat{\mathcal{X}}_k e^{j\measuredangle Y_k}$.

In the original MMSE STSA approach [1], the cost function is chosen as $C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2$. Recently, the MMSE-STSA estimator was generalized [2] by modifying the cost function as:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta})^2 \tag{2}$$

where the exponent β is a real parameter whose purpose is to control the associated estimator gain function and, consequently, the tradeoff between speech distortion and noise reduction. The case $\beta > 0$ was analyzed in [2] while the analysis was extended to the case $-2 < \beta < 0$ in [5]. We will refer to this estimator as the β-SA estimator.

In [3], the following weighted form of the MMSE STSA cost function was proposed:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k - \hat{\mathcal{X}}_k}{\mathcal{X}_k^{p}} \right)^2 \tag{3}$$

This work was supported by the *Fonds québécois de la recherche sur la nature et les technologies* and a grant from the *Natural Sciences and Engineering Research Council of Canada*.

where $p < 1$ is a real parameter[1]. This estimator is termed the WE estimator and was motivated by the masking properties of the ear. In fact, for $p > 0$, it forces a better clean speech estimation in regions where the STSA is smaller, and therefore less likely to mask noise remaining in the clean speech estimation. As for β in the β-SA estimator, p was also found to control the trade-off between speech distortion and noise reduction.

3. WEIGHTED β-SA ESTIMATOR

In this work, we seek to combine the β-SA and WE cost functions into a single cost function to take advantage of both perceptual interpretations that can be given to the parameters β and p (which will be further discussed in the next section). The proposed cost function is therefore:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta}}{\mathcal{X}_k^{p}} \right)^2 \tag{4}$$

where β and p are real parameters.

To obtain the Bayesian estimator corresponding to the cost function in (4), we need to minimize its expectation with respect to $\hat{\mathcal{X}}_k$. By doing so, we obtain:

$$\hat{\mathcal{X}}_k = \left(\frac{E\left\{ \mathcal{X}_k^{\beta-2p} | Y_k \right\}}{E\left\{ \mathcal{X}_k^{-2p} | Y_k \right\}} \right)^{\frac{1}{\beta}}. \tag{5}$$

Using the Gaussian statistical model in [1] where the complex spectrums (i.e. the Fourier expansion coefficients) of the clean speech and noise were considered to be independent, identically distributed Gaussian random variables with zero mean and variances $\sigma_x^2 = E\{\mathcal{X}_k^2\}$ and $\sigma_n^2 = E\{|N_k|^2\}$, respectively, we know (see [4] and Appendix A in [3]) that:

$$E\left\{ \mathcal{X}_k^m | Y_k \right\} = \lambda_k^{m/2} \Gamma\left(\frac{m}{2} + 1\right) M\left(-\frac{m}{2}, 1; -\upsilon_k\right) \tag{6}$$

where $\Gamma(x)$ is the gamma function, $M(a, b; z)$ is the confluent hypergeometric function and $m > -2$. Moreover,

$$\frac{1}{\lambda_k} = \frac{1}{E\{|N_k|^2\}} + \frac{1}{E\{\mathcal{X}_k{}^2\}}$$

and

$$\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \quad \xi_k = \frac{E\{\mathcal{X}_k{}^2\}}{E\{|N_k|^2\}}, \quad \gamma_k = \frac{|Y_k|^2}{E\{|N_k|^2\}},$$

where ξ_k acts as a long term estimator of the SNR and is called the *a priori* SNR while $\gamma_k - 1$ can be interpreted as the instantaneous SNR.

Using (6) in (5) with the appropriate values of the parameter m (i.e. $m = \beta - 2p$ for the numerator and $m = -2p$ for the denominator) , we can show that:

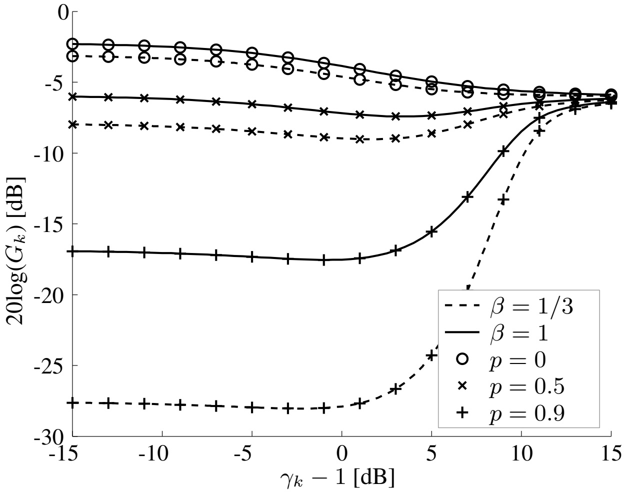
$$\hat{\mathcal{X}}_k = G_k |Y_k|$$

where

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left(\frac{\Gamma\left(\frac{\beta-2p}{2}+1\right) M\left(-\frac{\beta-2p}{2}, 1; -\upsilon_k\right)}{\Gamma\left(-p+1\right) M\left(p, 1; -\upsilon_k\right)} \right)^{1/\beta} \tag{7}$$

and $\beta > 2(p-1)$, $p < 1$. We will denote this estimator as the Weighted β-SA estimator (Wβ-SA).

The Wβ-SA estimator gain, G_k, depends on the parameters of the cost function (i.e. β and p) as well as on γ_k and ξ_k. Figure 1 presents gain curves as a function of the instantaneous SNR, $\gamma_k - 1$, for a fixed ξ_k = 0 dB and several β and p values. As can be observed, the estimator's gain decreases when p increases and increases when β increases. It is worth noting that a decrease in the gain will result in more noise reduction but will invariably introduce more speech distortion. Also, since the proposed estimator generalizes both the β-SA and WE estimators, the gains of the later can be obtained by setting $p = 0$ (for β-SA) and $\beta = 1$ (for WE).



(a)

Fig. 1. Estimator gain ($20\log(G_k)$) versus instantaneous SNR ($\gamma_k - 1$) for several β and p values (ξ_k = 0 dB).

It was shown in [3] that the WE estimator tends to a Wiener estimator as the instantaneous SNR tends to infinity. In fact, the more general Wβ-SA estimator also tends to a Wiener filter: we know from (13.1.5) in [6] that as $\gamma_k - 1 \longrightarrow \infty$, the confluent hypergeometric function $M(-\frac{m}{2}, 1; -\upsilon_k)$ can be written as:

$$M(-\frac{m}{2}, 1; -\upsilon_k) = \frac{\upsilon_k^{m/2}}{\Gamma(\frac{m}{2}+1)}. \tag{8}$$

Using (8) in (7) with the appropriate values of the parameter m, we have:

$$G_k = \frac{\xi_k}{1+\xi_k} \tag{9}$$

which is a Wiener filter gain.

4. PERCEPTUALLY RELEVANT β AND p VALUES

In this section, we will consider the choice of β and p values in the Wβ-SA estimator according to perceptual considerations.

Let us first consider the choice of the β value. Power laws have been used in the past to model the nonlinear relation between the intensity of sound and its perceived loudness [7]. Since loudness is more perceptually relevant than the sound's intensity, a cost function which would consider the difference in terms of the perceived loudness would be preferable to cost functions which consider the

[1] In [3], the equivalent cost function was proposed as $C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \mathcal{X}_k^p(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2$, but the form (3) was found more convenient here.

difference in terms of the sound intensity. An exponent of 1/3 (i.e. cubic root) has been proposed in [8] and used in [7] to perform the nonlinear transformation between intensity and perceived loudness, we therefore propose to set $\beta = 1/3$.

As can be observed from Figure 1, using $\beta = 1/3$ (as opposed to keeping $\beta = 1$) will imply a lower gain G_k which should therefore produce more noise reduction but will, however, also introduce more speech distortion.

We now look at the choice of the p value. The motivation for deriving the WE estimator was to favor a more accurate estimation of smaller STSA since they are less likely to mask noise remaining in the clean speech estimation. This was done by increasing the weight of smaller STSA in the cost function. Since the first formants, which contain most of the speech energy, are located at lower frequencies, higher frequencies should contain mainly small STSA. Therefore, it would be relevant to further increase the weights of the smaller STSA in the cost function for higher frequencies. This can be done by increasing p for higher frequencies. We therefore propose to modify the values of p as a function of frequency, i.e. p_k, increasing its value for higher frequencies.

We need to choose appropriately the values of p_k for each frequencies. In [3], the value of $p = 0.5$ has been suggested as a good compromise between the desired noise reduction performed by the estimator and the speech distortion introduced. This value can therefore be also regarded as being a good compromise between increasing the weight of smaller STSA while keeping an appropriate estimation error for higher STSA. Since the main part of the speech energy, which will contain most of the higher STSA, is approximately located below 2000 Hz (which includes most of the first two formants), we will keep this value up to 2000 Hz. For higher frequencies, we want to further increase the weights of smaller STSA. Since the total speech energy decreases as frequency increases, we therefore propose to linearly increase the value of p as a function of the frequency. Since the Wβ-SA estimator restricts p to $p < 1$, we will choose the highest value, i.e. $p = 0.9$, for the highest frequency. Therefore p_k will be given by:

$$p_k = \begin{cases} p_{low} & \text{if } f_k \leq 2000Hz \\ \frac{(f_k - 2000)(p_{high} - p_{low})}{F_s/2 - 2000} + p_{low} & \text{else} \end{cases} \tag{10}$$

where $p_{low} = 0.5$, $p_{high} = 0.9$, f_k is the frequency in Hz corresponding to spectral component k and F_s is the sampling frequency set to 16 kHz.

5. RESULTS

In this section, we compare the proposed estimator to the MMSE STSA, the MMSE log-STSA (LSA) [4] (where the cost function takes the form $C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\log \mathcal{X}_k - \log \hat{\mathcal{X}}_k)^2$) and the WE estimator (with $p = 0.5$ as suggested in [3]). Noisy speech signals were created according to ITU-T standard P.56 [9]. We present results using white noise and aircraft cockpit (buccaneer-1) noise [10]. All speech signals were sampled at 16 kHz and a raised-cosine window was used (512 samples, 32ms). A 75% overlap was used in the overlap-add synthesis method. All estimators used the *decision-directed* approach for the estimation of ξ_k [1] and a voice activity detector proposed in [11] was used to evaluate the noise spectral amplitude variance.

5.1. Objective results

We present objective results using the segmental SNR (SNR$_{seg}$) measure [12] over 30 Harvard sentences [13] (3 males, 3 females, 5 sentences each). To evaluate the relevance of the proposed values for β and p, i.e. $\beta = 1/3$ and $p = p_k$ (10), we present results using the Wβ-SA estimators for values of ($\beta = 1/3, p = 0.5$), ($\beta = 1, p = p_k$) and ($\beta = 1/3, p = p_k$). Note that the WE estimator corresponds to values of $\beta = 1$ and $p = 0.5$.

Table 1 shows the SNR$_{seg}$ results for SNRs of 0, 5 and 10 dB for white noise while Table 2 presents results for the aircraft cockpit noise.

Table 1. SNR$_{seg}$ values for MMSE STSA, LSA, WE and Wβ-SA estimators with white noise.

		0 dB	5 dB	10 dB
MMSE STSA [1]		0.47	3.92	7.14
LSA [4]		2.06	5.12	7.92
WE ($p = 0.5$) [3]		2.97	5.75	8.29
Wβ-SA	$\beta = 1/3, p = 0.5$	3.28	5.96	8.38
	$\beta = 1, p = p_k$	3.35	6.02	8.46
	$\beta = 1/3, p = p_k$	3.62	6.21	8.52

Table 2. SNR$_{seg}$ values for MMSE STSA, LSA, WE and Wβ-SA estimators with aircraft cockpit noise.

		0 dB	5 dB	10 dB
MMSE STSA [1]		-0.57	2.71	5.55
LSA [4]		0.35	3.41	5.93
WE ($p = 0.5$) [3]		0.79	3.63	5.99
Wβ-SA	$\beta = 1/3, p = 0.5$	0.90	3.67	5.94
	$\beta = 1, p = p_k$	0.83	3.65	5.99
	$\beta = 1/3, p = p_k$	0.94	3.68	5.93

As can be observed, setting $\beta = 1/3$ produces better results in terms of SNR$_{seg}$ than $\beta = 1$. Furthermore, modifying the value of p as a function of the frequency also produces better results than keeping it constant at $p = 0.5$ for all frequencies. The best results are therefore obtained by setting $\beta = 1/3$ and $p = p_k$. One exception to those observations is the aircraft cockpit noise at 10 dB where the best performance is obtained equally by the Wβ-SA ($\beta = 1, p = p_k$) and the WE estimators.

5.2. Subjective results

As a subjective measure, we used the MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) method (ITU-R Recommendation BS.1534-1 [14]) as implemented in [15]. In MUSHRA, the subjects are provided with the test utterances plus one reference and one hidden anchor and are asked to rate the different signals on a scale of 0 to 100, 100 being the best score. As the hidden anchor, we used a signal having an SNR of 5 dB less than the noisy signal to be enhanced. The listeners were allowed to listen to each sentence several times and always had access to the clean signal reference.

A total of 8 listeners (7 males, 1 female aged in the mid 20's to low 30's) participated in the test. A subset of two sentences (one male speaker, one female speaker) were chosen randomly from the thirty sentences used previously for the objective evaluation (the two same sentences for all subjects). Tests were performed in an isolated acoustic room using beyerdynamic DT880 headphones. In order to limit the length of the listening test, only the 0 dB case was considered. The average duration of a test was approximately 30 minutes per subject.

Table 3 presents the MUSHRA comparative results for the MMSE STSA, LSA and WE estimators along with those of the Wβ-SA estimator with the proposed values of $\beta = 1/3$ and $p = p_k$. As can be observed, the sentences enhanced using the Wβ-SA estimator were rated higher than those enhanced by the other estimators for both white and cockpit noises. Two-tailed paired t-tests revealed the advantage of the Wβ-SA estimator (i.e. the differences between the scores in Table 3) to be statistically significant within a 95% confidence interval.

Table 3. MUSHRA values for MMSE STSA, LSA, WE and Wβ-SA estimators with white and aircraft cockpit noise at 0 dB.

	White (0 dB)	Cockpit (0 dB)
MMSE STSA [1]	22.3	27.3
LSA [4]	33.8	37.1
WE [3]	42.4	47.6
Wβ-SA ($\beta = 1/3, p = p_k$)	56.8	55.3

5.3. Discussion

We can observe from Figure 1 that decreasing β as well as increasing p in the Wβ-SA estimator both result in a decrease in the corresponding gain G_k. Therefore, the gain of the Wβ-SA estimator with the proposed $\beta = 1/3$ and $p = p_k$ values is smaller than the gain of the WE estimator with $p = 0.5$ (which corresponds to the Wβ-SA with $\beta = 1$ and $p = 0.5$). The proposed estimator therefore produces more noise reduction than the WE estimator ($p = 0.5$) but also generates more speech distortion. Since the main speech energy is concentrated at lower frequencies, the frequency dependence of p allows to limit the speech distortion at lower frequencies and increase the noise reduction at higher frequencies.

On the one hand, due to the increased noise reduction at high frequencies, the proposed estimator should be more advantageous when the noise has more high frequency components. In fact, the proposed estimator showed more improvements both in terms of SNR$_{seg}$ and MUSHRA for the white noise than for the aircraft cockpit noise which has less high frequency components than white noise. On the other hand, by decreasing the gain for the higher frequencies, the high frequency components of speech, such as fricatives, will be distorted. This distortion will be hardly noticeable when the SNR of the noisy sentence is low and therefore the enhancement performed by the proposed estimator will be perceived as an improvement over that of the other tested estimators, however, the speech distortion could become noticeable when the SNR of the noisy sentence increases. In fact, the proposed estimator yielded more improvements over the other tested estimators in terms of SNR$_{seg}$ at lower SNRs than at higher SNRs.

6. CONCLUSION

In summary we proposed a new family of estimators for speech enhancement, the Wβ-SA. We showed that improvements could be achieved with respect to existing Bayesian estimators such as the MMSE STSA, LSA and WE estimators when choosing the parameter values of the Wβ-SA estimator (i.e. β and p) to account for the perceived loudness of sound and take advantage of the masking properties of the ear. In fact, the proposed Wβ-SA estimator showed better overall performances both in terms of SNR$_{seg}$ results and MUSHRA scores when considering white and aircraft cockpit noises.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[2] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, July 2005.

[3] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[5] E. Plourde and B. Champagne, "Further analysis of the β-order MMSE STSA estimator for speech enhancement," in *Proc. 20th IEEE Canadian Conf. on Electrical and Computer Eng.*, Vancouver, Canada, 2007, pp. 1594–1597.

[6] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Washington : U.S. Govt. Print. Off., 1964.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[8] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, pp. 153–181, 1957.

[9] ITU-T, "Recommendation P.56: Objective measurement of active speech level," 1993.

[10] Rice University, "Signal processing information base: Noise data," [Online] Available http://spib.rice.edu/spib/select_noise.html.

[11] J. Sohn and N. S. Kim, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[12] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. 5th Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998, pp. 2819–2822.

[13] IEEE Standards Publication No. 297, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, no. 3, Sept. 1969.

[14] ITU-R, "Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," 2001.

[15] E. Vincent, "MUSHRAM: A MATLAB interface for MUSHRA listening tests," [Online] Available http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/.